Estimating Effect Sizes From Graphs Using Electronic Calipers: An Analysis of the Reliability and Accuracy of a Data-Gathering Method

Amanda C. Barnard
Ross R. Vickers Jr.

Naval Health Research Center

Document No. 10-11

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. Approved for public release; distribution is unlimited.

.

This research has been conducted in compliance with all applicable federal regulations governing the protection of human subjects in research.

Naval Health Research Center
140 Sylvester Rd.
San Diego, California 92106-3521

Estimating Effect Sizes From Graphs Using Electronic Calipers: An Analysis of the Reliability and Accuracy of a Data-Gathering Method

Amanda C. Barnard
Ross R. Vickers Jr.

Warfighter Performance Department
Naval Health Research Center
140 Sylvester Road
San Diego, CA 92106-3521

e-mail: ross.vickers@med.navy.mil

This represents report 10-11 and was supported by the U.S. Department of Defense under work unit 60704. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. Approved for public release; distribution is unlimited. This research has been conducted in compliance with all applicable federal regulations governing the protection of human subjects in research.

Abstract

Valuable data that would strengthen meta-analyses are often presented in graphs without reported means and standard deviations. The true state of knowledge about investigative questions is not accurately represented because this data is not included in the analysis. This paper describes and evaluates a method for extracting estimated effect sizes from graphic presentations. Two studies were conducted to assess the reliability and accuracy of using electronic calipers to estimate effect sizes from bar and line graphs. The first study looked at the reliability of effect size estimates derived from measurements taken from published graphs showing changes in one-repetition maximum strength. The second study assessed the accuracy of effect size estimates computed from graphs that were constructed from known means and standard deviations. The first study demonstrated very high levels of test-retest and inter-rater reliability for the effect size estimates. The second study showed a close correspondence between the effect sizes estimated from the graphs and the known effect sizes used to construct the graphs. Thus, using electronic calipers to estimate effect sizes from graphs produces results that are accurate and reliable. Meta-analysts can confidently use this methodology to include results that have been presented only graphically.

Introduction

Meta-analysis is a tool for providing a systematic, quantitative summary of findings in a research domain. Meta-analysis integrates results from different studies by pooling effect sizes (Cooper & Hedges, 1994; Hedges & Olkin, 1985). This method has been used to summarize the state of the art in research in several areas of applied physiology (Lemura, von Duvillard, & Mookerjee, 2000; Londeree, 1997; Rhea, Alvar, Burkett, & Ball, 2003).

Standard practices in reporting applied physiological research findings pose some problems for meta-analysis. The use of graphs is one such problem. To illustrate, consider the experimental-control group research design that is typically used in applied physiology studies. Measurements are taken on both groups before and after introducing an intervention. For example, resistance training program studies commonly involve assigning study participants to treatment and control groups and obtaining strength measurements before and after the training period for the experimental group. If the study findings give the means and standard deviations of the pre and posttraining measurements, an effect size suitable for use in a meta-analysis can be computed directly from the information given. The meta-analysis problems arise when the results are reported only in graphs.

Effect sizes cannot be computed directly from graphs—so the meta-analyst must either ignore those studies, or use the graphs to estimate the appropriate statistics. Ignoring the studies reduces the body of evidence available to describe the state of the art in the area of interest. This point is important because graphic presentation forms a significant proportion of the available evidence for some reviews (e.g., Galvao and Taaffe, 2004). In such cases, omitting the results embodied in the graphical representations certainly reduces the precision of aggregate effect size estimates and may also bias the aggregate estimates.

A method of estimating effect sizes from graphs would be a useful tool to ensure that meta-analyses accurately represent the state of knowledge within a research domain. This report evaluates a simple method of converting graphs to estimates of means and standard deviations (SDs). The method relies on caliper measurements taken by one or more investigators. Two questions about estimating effect sizes from graphs were addressed: (1) How reliable are the measurements? (2) How accurate are the estimated effect sizes derived from the measurements? Two studies were conducted to address these questions. The first study assessed the test-retest and inter-rater reliability of effect size estimates derived from published graphs. The second study looked at the accuracy of the measurements by comparing the graphic effect sizes (computed from the measurements) with the tabular effect sizes (computed from the reported means and SDs). Taken together, the studies demonstrate that the conversion of graphs to effect size estimates is reliable and accurate.

Study One: Reliability Testing

Methods

The authors of this report selected the graphs and conducted the data preparation, measuring, coding, and analysis.

Forty-one graphs presenting one-repetition maximum (1RM) strength measurements from studies that employed pre-test/post-test designs to assess resistance training program effectiveness were selected for analysis from four journals: Journal of Strength and Conditioning Resistance, Medicine in Science and Sports Exercise, European Journal of Applied Physiology, and Acta Physiology Scandinavica. The graphs were taken from 14 studies that presented 1RM data in graphic form only. The articles are indicated in the reference section.

A maximum of six pre and posttest measurements were taken from each study to ensure adequate study representation. Differences in the graph generation between studies were expected to have negligible effects on the measurement accuracy from the graphs. The graphs were extracted from a variety of sources to ensure that the findings had general applicability. Each measurement was taken from either a bar or line graph that presented 1RM means and SDs/errors. No control group measurements were taken. Furthermore, when results were presented for intermediate outcomes during training, analyses were limited to the initial and final measurements.

Once the graphs were selected, the following steps were taken to obtain individual measurements:

Step 1. Converted graphs to a workable size while maintaining original proportions (this can be done on most photocopiers, or in Microsoft Word or PowerPoint programs).
Step 2. Used the 797B-8/200 electronic caliper (Starrett, Athol, MA).
Step 3. Turned on the calipers, set them to measure in millimeters (mm), and zeroed them out.
Step 4. Placed the bottom point of the calipers at the center of the baseline (x-axis); then opened the calipers so the top point was near the center of the bar or line being measured.
Step 5. Carefully adjusted the caliper until the top point was placed precisely at the center of the line being measured.
Step 6. Held calipers at approximately a 45 degree angle away from the investigator to ensure visibility of the graphs when taking measurements.
Step 7. Recorded the digital reading of the distance to the nearest 100th of a mm.[1]

The following steps were taken to compute the estimated effect sizes for bar graphs, line graphs, and broken line graphs:

[1] The calipers are certified to be accurate within .02mm for measurements ranging from 0-100mm and within .03 mm for measurements ranging from 100-300mm.

1. The bar graph below is used to illustrate the following steps (Figure 1):
   Step 1. Measured and recorded mm on y-axis from the baseline (x-axis) to the maximum value (kg) (A).
   Step 2. Created scale conversion unit by dividing the kgs on the y-axis (100) by the y-axis mm.
   Step 3. Measured and recorded the total mm from the center of the baseline (x-axis) to the top of the error bars (B).
   Step 4. Measured and recorded the total mm from the center of the baseline (x-axis) to the top of the bar representing the mean (C).
   Step 5. Converted to kgs using the conversion unit to arrive at the estimated mean.
   Step 6. Calculated the difference between the measurement to the top of the error bars and the measurement for the mean to get the value representing the SD/error (B-C).
   Step 7. Converted value from Step 6 to kgs using the conversion unit to arrive at the estimated SD/error.
   Step 8. Calculated effect size using estimated means and SDs.

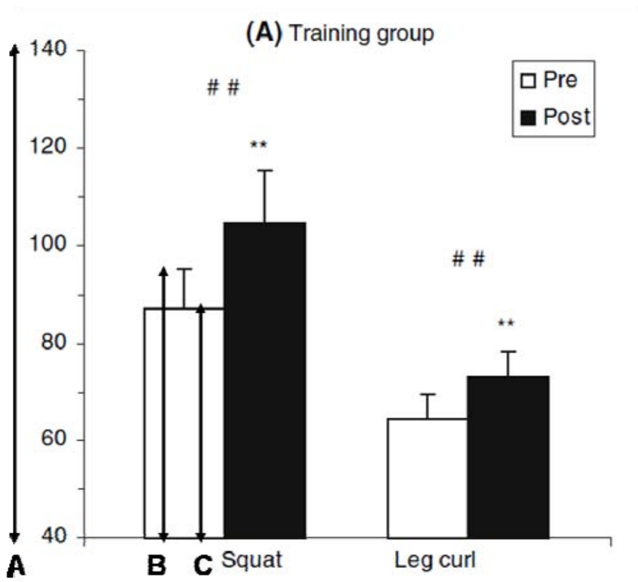

Figure 1. Hansen, Raastad, and Hallen (2007)

2. The line graph below is used to illustrate the following steps (Figure 2).
   Step 1. Measured and recorded mm on y-axis from the baseline (x-axis) to the maximum value (kg) (A).
   Step 2. Created scale conversion unit by dividing the kgs on the y-axis (40) by the y-axis mm.
   Step 3. Measured and recorded the total mm from the center of the baseline (x-axis) to the top of the error bars (B).
   Step 4. Measured and recorded the total mm from the center of the baseline (x-axis) to the center of the line representing the mean (C)
   Step 5. Converted to kgs using the conversion unit to arrive at the estimated mean.

Step 6. Calculated the difference (absolute value) between the measurement for the error bars and the measurement for the mean to get the value representing the SD/error (B-C).
Step 7. Converted value from Step 6 to kgs using the conversion unit to arrive at the estimated SD/error.
Step 8. Calculated effect size using estimated means and SDs.

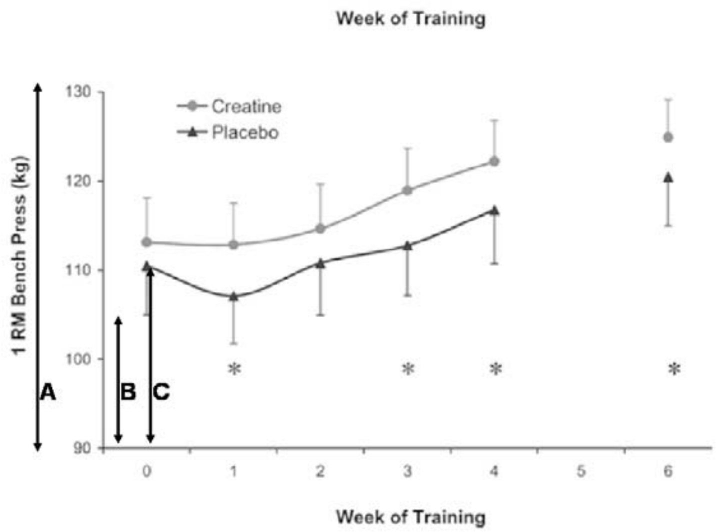

Figure 2. Volek et al. (2004)

3. The broken line graph below is used to illustrate the following steps (Figure 3) (elderly data)[2]:
   Step 1. Recorded mm from the baseline to the beginning of the continuous part of the scale (A).
   Step 2. This distance became the line graph adjustment value.
   Step 3. Created scale conversion unit by dividing the kgs (20) on the y-axis by the y-axis mm (B).
   Step 4. Took all subsequent measurements (means and SDs/errors) following the same procedures as the line graphs without breaks in the y-axis.
   Step 5. Once all data was recorded, subtracted the line graph adjustment value from the unadjusted means and SD/error measurements to get the actual mean and SD/error values. For example, to get the baseline value for the elderly, mean = (C-A) and SD = (D-[C-A]).
   Step 6. Converted values from Step 4 to kgs using the conversion unit to arrive at the estimated mean and standard deviation/error.
   Step 7. Calculated effect size using estimated means and SDs.

[2] The break in the y-axis is represented by * in Figure 3.

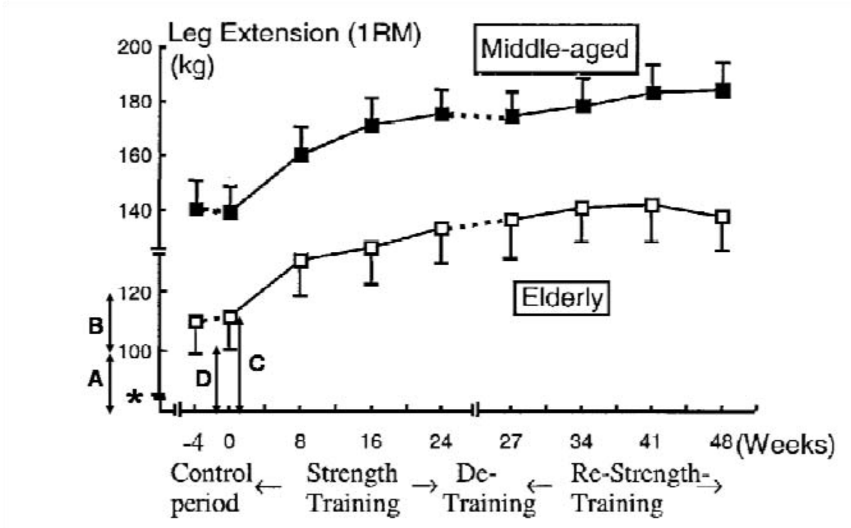

Figure 3. Hakkinen, Alen, Kallinen, Newton, and Kraemer (2000)

Two investigators each took two measurements—which yielded four estimates for each effect size. This provided data used to assess intra- and inter-rater reliability.

Statistical data analysis.

Analysis of the data was conducted using the computer program SPSS-PC (Version 17). For part one, a reliability analysis was conducted to determine intra- and inter-rater correlations.

Results

Study one's means, SDs, and intra- and inter-rater correlations for the effect sizes are presented in Table 1.

Table 1

Intra and Inter-Rater Correlations for Study One

Rater	Trial	Mean	SD	Rater1 Trial 1	Rater 1 Trial 2	Rater 2 Trial 1	Rater 2 Trial 2
1	1	1.36	.95	1.000	.997	.996	.996
1	2	1.38	.97	.997	1.000	.993	.995
2	1	1.35	.96	.996	.993	1.000	.994
2	2	1.35	.90	.996	.995	.994	1.000

The results of study one show an average correlation of 0.997 between estimated effect sizes. Higher correlations were seen between test-retest than inter-rater measurements, but all correlations were strong (≥0.993). Furthermore, no difference was seen between the reliability of bar graph and line graph measurements (both correlations = 0.997).

Discussion

Study one demonstrated that using electronic calipers to take measures from graphs of the effects of strength training programs produced results that were highly reliable. The evidence demonstrated strong test-retest reliability and inter-rater reliability. However, it was still unknown how accurate these graphic effect size estimates were when compared to the effect size estimates that would have been obtained if the tabular means and standard deviations for the effects had been reported. Study two addressed the question of how accurate the estimates of effect sizes are when derived from graph measurements.

Study Two: Accuracy Testing

Methods

Two researchers each selected a set of 20 paired pretraining and posttraining 1RM measurements at random from a set of 728 effect sizes reported in 181 articles containing data on the effects of resistance training on maximal strength. The means and SDs for the measurements needed to construct the effect sizes were extracted. The articles that provided those statistics are indicated in the reference section.

The information extracted from the articles was used to construct 40 effect size estimates as follows:

a. Each researcher used Microsoft Excel to create 20 bar graphs with error bars from the set of 20 pretraining and posttraining means and standard deviations/errors that he or she had selected.
b. Each investigator used the methods described in Study 1 to obtain the measurements needed to compute effect size estimates from the graphs created by the other investigator. Those measurements were taken without any knowledge of the true means and SDs/errors that had been used to create the graphs.
c. After measurements were recorded, effect sizes were calculated as described in study one.

A data file was constructed that included the effect size estimates derived from the means and standard deviations reported in the original articles (tabular effect sizes) and the effect size estimates derived from the graph measurements (graphic effect size estimates). The file contained 40 paired observations—20 from each investigator.

Statistical data analysis.

Paired *t* tests were conducted using SPSS-PC. The *t* tests compared effect size computed from the reported means and standard deviations to the graph-based estimates.

Results

Means, SDs correlation, *t* values, and significance for study two’s effect sizes are reported in Table 2.

Table 2

Effect Sizes for Study Two

	Graphic		Tabular					
	Mean	SD	Mean	SD	Effect Size[3]	Correlation	*t*	Sig
Rater 1	.80	.75	.79	.74	0.008	0.9998	3.26	.004
Rater 2	.96	.55	.95	.54	0.007	0.9996	2.18	.042

The correlations between the graphic and the tabular effect sizes were 0.9998 and 0.9996 for raters 1 and 2 respectively. There appears to be a statistically significant difference between the graphic and tabular effect sizes (rater 1, p=0.004; rater 2, p=0.042).

The statistically significant differences were due only to the high correlation of the graphic effect sizes with the tabular effect sizes. To demonstrate this, the differences between tabular and graphic means were converted to effect sizes using the following equation: Effect Size = (Graphic Mean-Tabular Mean)/Tabular SD. When the differences between the tabular and graphic effect sizes were converted to effect sizes using the means and SDs (see Table 2), the results were 0.008 for rater 1 and 0.007 for rater 2. These effect sizes would have to be more than 10 times larger to be considered practically or theoretically important (Cohen, 1988).

Discussion

Graphic measurements yield accurate effect size estimates. As a consequence, effect size estimates derived from graphs can be combined with those derived from reported means and SDs. In this study, the effect sizes derived by the two methods were almost perfectly correlated. In addition, the means and SDs were virtually identical. The effect sizes derived from graphs were smaller than the true values, but the difference was only .01 for each investigator. The difference was statistically significant in each case, but it was too small relative to the SD of the effect sizes to be of theoretical or practical importance.

General Discussion and Conclusions

Measurements taken from graphs can provide acceptable effect size estimates. The studies reported here demonstrated that those measurements yield reliable and accurate estimates of effect sizes. Test-retest and inter-rater reliability were high and the difference between known effect sizes and estimated effect sizes were trivial.

The results of these studies may have a significant limitation. The work reported here relied on electronic calipers with digital measurement readouts. Other calipers might produce somewhat lower reliability and accuracy if the raw measurements were less reliable.

Several issues arose when taking measurements of the graphic data. To assure accuracy in extracting data from graphs using electronic calipers, potential errors must be avoided. The following steps should be taken to avoid possible sources of error:

- Close calipers completely when zeroing out.
- Properly account for breaks in the scale on the y-axis in line graphs (see Figure 3).
- Make sure to measure from the numerical values, not the break lines, when working with breaks on the y-axis (see Figure 3).
- Be consistent in where the line measurements are being taken (e.g. in a bar graph if you place the bottom caliper point in the middle of the x-axis, place the top caliper point in the middle of the line on the top of the bar). This is particularly important if the lines are thick.
- Place the calipers on the actual baseline for all measurements in graphs where the bars extended below the baseline.

Electronic calipers appear to be a reliable and accurate tool for measuring and converting graphs to estimates of effect sizes. This study looked at both test-retest and inter-rater reliability and found high correlations among both. These findings demonstrate a novel method of incorporating a more comprehensive selection of the existing research when conducting meta-analyses.

References

*Studies used in reliability study (study 1)
**Studies used in accuracy study (study 2)

**Almasbakk, B., & Hoff, J. (1996). Coordination, the determinant of velocity specificity? *J Appl Physiol, 81*(5), 2046-2052.

**Berger, R. A. (1962). Effect of varied weight training programs on strength. *Res Q, 33*, 168-181.

*Bishop, D., Jenkins, D. G., Mackinnon, L. T., McEniery, M., & Carey, M. F. (1999). The effects of strength training on endurance performance and muscle characteristics. *Med Sci Sports Exerc, 31*(6), 886-891.

**Braith, R. W., Graves, J. E., Leggett, S. H., & Pollock, M. L. (1993). Effect of training on the relationship between maximal and submaximal strength. *Med Sci Sports Exerc*, *25*(1), 132-138.

*Brentano, M. A., Cadore, E. L., Da Silva, E. M., Ambrosini, A. B., Coertjens, M., Petkowicz, R., Viero, I., & Kruel, L. F. (2008). Physiological adaptations to strength and circuit training in postmenopausal women with bone loss. *J Strength Cond Res, 22*(6), 1816-1825.

**Buford, T. W., Rossi, S. J., Smith, D. B., & Warren, A. J. (2007). A comparison of periodization models during nine weeks with equated volume and intensity for strength. *J Strength Cond Res, 21*(4), 1245-1250.

*Candow, D. G., & Burke, D. G. (2007). Effect of short-term equal-volume resistance training with different workout frequency on muscle mass and strength in untrained men and women. *J Strength Cond Res, 21*(1), 204-207.

**Charette, S. L., McEvoy, L., Pyka, G., Snow-Harter, C., Guido, D., Wiswell R.A., & Marcus, R. (1991). Muscle hypertrophy response to resistance training in older women. *J Appl Physiol, 70*(5), 1912-1916.

*Chrusch, M. J., Chilibeck, P. D., Chad, K. E., Davison, K. S., & Burke, D. G. (2001). Creatine supplementation combined with resistance training in older men. *Med Sci Sports Exerc, 33*(12), 2111-2117.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (Second ed.).* Hillsdale, NJ: Erlbaum.

Cooper, H., & Hedges, L.V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

**Cortez-Cooper, M. Y., Anton, M. M., Devan, A. E., Neidre, D. B., Cook, J. N., & Tanaka, H. (2008). The effects of strength training on central arterial compliance in middle-aged and older adults. *Eur J Cardiovasc Prev Rehabil, 15*(2), 149-155.

**Cowley, P. M., Swensen, T., & Sforzo, G. A. (2007). Efficacy of instability resistance training. *Int J Sports Med, 28*(10), 829-835.

**DeBeliso, M., Harris, C., Spitzer-Gibson, T., & Adams, K. J. (2005). A comparison of periodised and fixed repetition training protocol on strength in older adults. *J Sci Med Sport, 8*(2), 190-199.

**Fatouros, I. G., Kambas, A., Katrabasas, I., Leontsini, D., Chatzinikolaou, A., Jamurtas, A.Z. ...Taxildaris, K. (2006). Resistance training and detraining effects on flexibility performance in the elderly are intensity-dependent. *J Strength Cond Res, 20*(3), 634-642.

**Fatouros, I. G., Tournis, S., Leontsini, D., Jamurtas, A. Z., Sxina, M., Thomakos, P. … Mitrakou, A. (2005). Leptin and adiponectin responses in overweight inactive elderly following resistance training and detraining are intensity related. *J Clin Endocrinol Metab, 90*(11), 5970-5977.

Galvao, D. A., & Taaffe, D. R. (2004). Single- vs. multiple-set resistance training: recent developments in the controversy. *J Strength Cond Res, 18*(3), 660-667.

**Galvao, D. A., & Taaffe, D. R. (2005). Resistance exercise dosage in older adults: single- versus multiset effects on physical performance and body composition. *J Am Geriatr Soc, 53*(12), 2090-2097.

*Hakkinen, K., Alen, M., Kallinen, M., Newton, R. U., & Kraemer, W. J. (2000). Neuromuscular adaptation during prolonged strength training, detraining and re-strength-training in middle-aged and elderly people. *Eur J Appl Physiol, 83*(1), 51-62.

*Hakkinen, K., Kraemer, W. J., Newton, R. U., & Alen, M. (2001). Changes in electromyographic activity, muscle fibre and force production characteristics during heavy resistance/power strength training in middle-aged and older men and women. *Acta Physiol Scand, 171*(1), 51-62.

*Hansen, E. A., Raastad, T., & Hallen, J. (2007). Strength training reduces freely chosen pedal rate during submaximal cycling. *Eur J Appl Physiol, 101*(4), 419-426.

*Hass, C. J., Garzarella, L., de Hoyos, D., & Pollock, M. L. (2000). Single versus multiple sets in long-term recreational weightlifters. *Med Sci Sports Exerc, 32*(1), 235-242.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

**Hoffman, J. R., Kraemer, W. J., Fry, A. C., Deschenes, M., & Kemp, M. (1990). The effects of self-selection for frequency of training in a Winter conditioning program for football. *Journal of Applied Sport Science Research, 4*(3), 76-82.

**Izquierdo, M., Hakkinen, K., Ibanez, J., Garrues, M., Anton, A., Zuniga, A. … Gorostiaga, M. (2001). Effects of strength training on muscle power and serum hormones in middle-aged and older men. *J Appl Physiol, 90*(4), 1497-1507.

**Kerksick, C. M., Rasmussen, C. J., Lancaster, S. L., Magu, B., Smith, P., Melton, C. … Kreider, R.B. (2006). The effects of protein and amino acid supplementation on performance and training adaptations during ten weeks of resistance training. *J Strength Cond Res, 20*(3), 643-653.

**Kerr, D., Morton, A., Dick, I., & Prince, R. (1996). Exercise effects on bone mass in postmenopausal women are site-specific and load-dependent. *J Bone Miner Res, 11*(2), 218-225.

**Kirk, E. P., Washburn, R. A., Bailey, B. W., LeCheminant, J. D., & Donnelly, J. E. (2007). Six months of supervised high-intensity low-volume resistance training improves strength independent of changes in muscle mass in young overweight men. *J Strength Cond Res, 21*(1), 151-156.

**Kotzamanidis, C., Chatzopoulos, D., Michailidis, C., Papaiakovou, G., & Patikas, D. (2005). The effect of a combined high-intensity strength and speed training program on the running and jumping ability of soccer players. *J Strength Cond Res, 19*(2), 369-375.

**Kraemer, W. J. (1997). A series of studies: The physiological basis for strength training in American football: Fact over philosophy. *Journal of Strength and Conditioning Research, 11*, 131-142.

*Kraemer, W. J., Mazzetti, S. A., Nindl, B. C., Gotshalk, L. A., Volek, J. S., Bush, J. A. … Hakkinen, K. (2001). Effect of resistance training on women's strength/power and occupational performances. *Med Sci Sports Exerc, 33*(6), 1011-1025.

**Kraemer, W. J., Ratamess, N. A., Volek, J. S., Hakkinen, K., Rubin, M. R., French, D. N. … Dioguardi, F.S. (2006). The effects of amino acid supplementation on hormonal responses to resistance training overreaching. *Metabolism, 55*(3), 282-291.

Lemura, L. M., von Duvillard, S. P., & Mookerjee, S. (2000). The effects of physical training of functional capacity in adults. Ages 46 to 90: A meta-analysis. *J Sports Med Phys Fitness, 40*(1), 1-10.

Londeree, B. R. (1997). Effect of training on lactate/ventilatory thresholds: A meta-analysis. *Med Sci Sports Exerc, 29*(6), 837-843.

*Mangine, G. T., Ratamess, N. A., Hoffman, J. R., Faigenbaum, A. D., Kang, J., & Chilakos, A. (2008). The effects of combined ballistic and heavy resistance training on maximal lower- and upper-body strength in recreationally trained men. *J Strength Cond Res*, *22*(1), 132-139.

**Marcinik, E. J., Potts, J., Schlabach, G., Will, S., Dawson, P., & Hurley, B. F. (1991). Effects of strength training on lactate threshold and endurance performance. *Med Sci Sports Exerc, 23*(6), 739-743.

*Mazzetti, S. A., Kraemer, W. J., Volek, J. S., Duncan, N. D., Ratamess, N. A., Gomez, A. L. … Fleck, S.J. (2000). The influence of direct supervision of resistance training on strength performance. *Med Sci Sports Exerc, 32*(6), 1175-1184.

**Moss, B. M., Refsnes, P. E., Abildgaard, A., Nicolaysen, K., & Jensen, J. (1997). Effects of maximal effort strength training with different loads on dynamic strength, cross-sectional area, load-power and load-velocity relationships. *Eur J Appl Physiol Occup Physiol, 75*(3), 193-199.

**Pipes, T. V. (1977). Physiological responses of fire fighting recruits to high intensity training. *J Occup Med, 19*(2), 129-132.

Rhea, M. R., Alvar, B. A., Burkett, L. N., & Ball, S. D. (2003). A meta-analysis to determine the dose response for strength development. *Medicine and Science in Sports and Exercise, 35*(3), 456-464.

**Rhea, M. R., Ball, S. D., Phillips, W. T., & Burkett, L. N. (2002). A comparison of linear and daily undulating periodized programs with equated volume and intensity for strength. *J Strength Cond Res*, *16*(2), 250-255.

**Schlicht, J., Camaione, D. N., & Owen, S. V. (2001). Effect of intense strength training on standing balance, walkin speed, and sit-to-stand performance in older adults. *J Gerontol A Biol Sci Med Sci, 56*, M281-M286.

**Schlumberger, A., Stec, J., & Schmidtbleicher, D. (2001). Single- vs. multiple-set strength training in women. *J Strength Cond Res, 15*(3), 284-289.

**Tarnopolsky, M., Zimmer, A., Paikin, J., Safdar, A., Aboud, A., Pearce, E. … Doherty, T. (2007). Creatine monohydrate and conjugated linoleic acid improve strength and body composition following resistance exercise in older adults. *PLoS ONE, 2*(10), e991.

**Terzis, G., Georgiadis, G., Stratakos, G., Vogiatzis, I., Kavouras, S., Manta, P. … Blomstrand, E. (2008). Resistance exercise-induced increase in muscle mass correlates with p70S6 kinase phosphorylation in human subjects. *Eur J Appl Physiol, 102*(2), 145-152.

**Tracy, B. L., & Enoka, R. M. (2006). Steadiness training with light loads in the knee extensors of elderly adults. *Med Sci Sports Exerc, 38*(4), 735-745.

**Trappe, S., Williamson, D., & Godard, M. (2002). Maintenance of whole muscle strength and size following resistance training in older men. *J Gerontol A Biol Sci Med Sci, 57*(4), B138-143.

**Vandenberghe, K., Goris, M., Van Hecke, P., Van Leemputte, M., Vangerven, L., & Hespel, P. (1997). Long-term creatine intake is beneficial to muscle performance during resistance training. *Journal of Applied Physiology, 83*(6), 2055-2063.

**Villareal, D. T., & Holloszy, J. O. (2006). DHEA enhances effects of weight training on muscle mass and strength in elderly women and men. *Am J Physiol Endocrinol Metab, 291*(5), E1003-1008.

*Vissing, K., Brink, M., Lonbro, S., Sorensen, H., Overgaard, K., Danborg, K. … Aagaard, P. (2008). Muscle adaptations to plyometric vs. resistance training in untrained young men. *J Strength Cond Res, 22*(6), 1799-1810.

*Volek, J. S., Ratamess, N. A., Rubin, M. R., Gomez, A. L., French, D. N., McGuigan, M. M. … Kraemer, W.J. (2004). The effects of creatine supplementation on muscular performance and body composition responses to short-term resistance training overreaching. *Eur J Appl Physiol, 91*(5-6), 628-637.

**Westcott, W. L., Winett, R. A., Anderson, E. S., Wojcik, J. R., Loud, R. L., Cleggett, E., & Glover, S. (2001). Effects of regular and slow speed resistance training on muscle strength. *J Sports Med Phys Fitness, 41*(2), 154-158.

**Wilmore, J. H., Parr, R. B., Girandola, R. N., Ward, P., Vodak, P. A., Barstow, T. J. … Leslie, P. (1978). Physiological alterations consequent to circuit weight training. *Med Sci Sports, 10*(2), 79-84.

**Wilson, G. J., Murphy, A. J., & Walshe, A. (1996). The specificity of strength training: the effect of posture. *Eur J Appl Physiol Occup Physiol, 73*(3-4), 346-352.

*Winchester, J. B., McBride, J. M., Maher, M. A., Mikat, R. P., Allen, B. K., Kline, D. E., & McGuigan, M.R. (2008). Eight weeks of ballistic exercise improves power independently of changes in strength and muscle fiber type expression. *J Strength Cond Res, 22*(6), 1728-1734.

**Woolstenhulme, M. T., Conlee, R. K., Drummond, M. J., Stites, A. W., & Parcell, A. C. (2006). Temporal response of desmin and dystrophin proteins to progressive resistance exercise in human skeletal muscle. *J Appl Physiol, 100*(6), 1876-1882.

REPORT DOCUMENTATION PAGE

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB Control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (DD MM YY)
02 02 10

2. REPORT TYPE
Technical report

3. DATES COVERED (from – to)
April 2009 – February 2010

4. TITLE
Estimating Effect Sizes From Graphs Using Electronic Calipers: An Analysis of the Reliability and Accuracy of a Data-Gathering Method

5a. Contract Number:
5b. Grant Number:
5c. Program Element Number:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number: 60704

6. AUTHORS
Barnard, Amanda C., Vickers, Ross R. Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Commanding Officer
Naval Health Research Center
140 Sylvester Rd
San Diego, CA 92106-3521

8. PERFORMING ORGANIZATION REPORT NUMBER
10-11

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

Commanding Officer	Commander
Naval Medical Research Center	Navy Medicine Support Command
503 Robert Grant Ave	P.O. Box 140
Silver Spring, MD 20910-7500	Jacksonville, FL 32212-0140

10. SPONSOR/MONITOR'S ACRONYM(S)
NMRC/NMSC

11. SPONSOR/MONITOR'S REPORT NUMBER(s)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Valuable data that would strengthen meta-analyses are often presented in graphs without reported means and standard deviations. This indicates potential data may not be used, and the true state of knowledge about the investigative question is not accurately represented. This paper describes and evaluates a method for extracting estimated effect sizes from graphic presentations. Two studies were conducted to assess the reliability and accuracy of using electronic calipers to estimate effect sizes from bar and line graphs. The first study looked at the reliability of effect size estimates derived from measurements taken from published graphs showing changes in one repetition maximum strength. The second study assessed the accuracy of effect size estimates computed from graphs that were constructed from known means and standard deviations. The first study demonstrated very high levels of test-retest and inter-rater reliability for the effect size estimates. The second study showed a close correspondence between the effect sizes estimated from the graphs and the known effect sizes used to construct the graphs. Thus, using electronic calipers to estimate effect sizes from graphs produces results that are accurate and reliable. Meta-analysts can confidently use this methodology to include results that have been presented only graphically.

15. SUBJECT TERMS
converting graphs, calipers, estimated effect sizes, meta-analysis, data gathering

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	18a. NAME OF RESPONSIBLE PERSON
UNCL	UNCL	UNCL	UNCL	15	Commanding Officer

18b. TELEPHONE NUMBER (INCLUDING AREA CODE)
COMM/DSN: (619) 553-8429

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18